Heritage Health Claims Data

October 2, 2020

1 Predicting Hospital Admissions using Claims Data

1.1 Summary

With the rise of readily available data, health-centers nationwide are actively working to minimize costs while producing optimal health outcomes. The only question remains is as follows: how do we predict costs? Hospitalization time is a key driver in healthcare billing. In the **Heritage Health Data Challenge** via Kaggle, I sought to address this via basic classification model building. In addition to exploratory data analysis and feature engineering, I fit three models. The **random forest** algorithm was the most accurate, yielding an accuracy rate of 86.7%

1.2 Introduction

Health informatics can carry significant impact with regards to costs and availability of services. The Heritage Health Competition was a past data competition hosted on Kaggle. Participants use available patient data to predict which patients are more likely to experience readmission.

In this project, I use the past datasets to conduct data cleaning, exploratory data analysis, modeling, and appropriate predictive analysis.

1.3 Data Processing

The datasets were released via Kaggle in CSV formats. They contain many instances of incomplete cases and require extensive cleaning. The tables were pulled from a relational database, in which the member id is the primary field linking tables. Therefore, joins are required; the **members** and **target** tables have one-to-one relationships, they can be merged using left and/or inner joins. The **drugs** and **labs** tables have a one-to-many relationship with the member table, as they contain records on a yearly basis.

1.3.1 Selecting Predictors

[11]:
```
   MemberID AgeAtFirstClaim  Sex  ClaimsTruncated  DaysInHospital Year  \
0       210           30-39  NaN              0.0             0.0   Y1
1       210           30-39  NaN              0.0             0.0   Y3
2      3197             0-9    F              0.0             0.0   Y1
3      3197             0-9    F              0.0             0.0   Y2
```

```
4      3197             0-9    F             0.0             0.0   Y3

  DrugCount LabCount  AMI  APPCHOL  …  RENAL2  RENAL3  RESPR4  ROAMI  \
0         2        0  0.0      0.0  …     0.0     0.0     0.0    0.0
1         2        0  0.0      0.0  …     0.0     0.0     0.0    0.0
2         1        0  0.0      0.0  …     0.0     0.0     1.0    0.0
3         2        0  0.0      0.0  …     0.0     0.0     1.0    0.0
4         1        0  0.0      0.0  …     0.0     0.0     1.0    0.0

   SEIZURE  SEPSIS  SKNAUT  STROKE  TRAUMA  UTI
0      0.0     0.0     0.0     0.0     0.0  0.0
1      0.0     0.0     0.0     0.0     0.0  0.0
2      0.0     0.0     0.0     0.0     0.0  0.0
3      0.0     0.0     0.0     0.0     0.0  0.0
4      0.0     0.0     0.0     0.0     0.0  0.0

[5 rows x 53 columns]
```

1.3.2 Outcome Variable

[12]:

```
              MemberID   Year   DSFS  PrimaryConditionGroup
LengthOfStay
1 day            25210  25210  22532                  24541
1- 2 weeks         358    358    322                    276
2 days            2767   2767   1944                   2445
2- 4 weeks         318    318    293                    289
26+ weeks            1      1      1                      1
3 days            1014   1014    755                    848
4 days             418    418    350                    312
4- 8 weeks         431    431    418                    414
5 days             155    155    129                    109
6 days              62     62     55                     28
```

[14]:

```
   MemberID Year         DSFS PrimaryConditionGroup LengthOfStay  \
0         4   Y2   0- 1 month                RESPR4          NaN
1       210   Y1   0- 1 month              GIOBSENT       2 days
3       210   Y1   0- 1 month                GYNEC1          NaN
4       210   Y1  1- 2 months                MSC2a3          NaN
6       210   Y1  3- 4 months                PRGNCY          NaN

   length_recoded
0             0.0
1             2.0
3             0.0
4             0.0
6             0.0
```

[15]:

	MemberID	Year	length_recoded
0	4	Y2	0.0
1	210	Y1	2.0
2	210	Y2	0.0
3	210	Y3	0.0
4	3197	Y1	0.0

1.3.3 Feature Engineering

One aspect of this project, which may differ from how other participants approached the challenge, entails my experience as a hospital volunteer, a public health student, and later a research assistant. Based on this, rather than employing forward or backward stepwise model building, I will be deliberately selecting features that have documented impacts on health.

One feature that I will be constructing is an SES categorical variable (`low_SES`), derived from the pay delay field. Pay delays can be the result of financial hardship, as I've learned through first hand experience. Socioeconomic status is a key determinant of health and will therefore be included in model building.

Another feature I will be adding is the count of timepoints within a year (`time_count`) in which a patient has a claim. So if a patient has a claim at 0-1 months and 3-4 months during Year One, this feature would be a value of 2.

[17]:

	MemberID	PayDelay
0	4	43
1	210	57
2	210	162+
3	210	151
4	210	22

[19]: (154212, 2)

[21]:

	MemberID	Year	DSFS
0	4	Y2	1
1	210	Y1	8
2	210	Y2	6
3	210	Y3	4
4	3197	Y1	5

1.3.4 Merging Datasets Back Together

[23]:

	MemberID	AgeAtFirstClaim	Sex	ClaimsTruncated	DaysInHospital	Year	\
0	210	30-39	NaN	0.0	0.0	1	
1	210	30-39	NaN	0.0	0.0	3	
2	3197	0-9	F	0.0	0.0	1	
3	3197	0-9	F	0.0	0.0	2	

```
4      3197           0-9    F          0.0          0.0    3
5      3713         40-49    F          0.0          0.0    2
6      3741         70-79    F          0.0          0.0    2
7      3889           NaN    F          0.0          0.0    1
8      4048         50-59    M          0.0          0.0    3
9      4187         50-59    F          0.0          0.0    1

   DrugCount  LabCount  AMI  APPCHOL  …  ROAMI  SEIZURE  SEPSIS  SKNAUT  \
0        2.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
1        2.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
2        1.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
3        2.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
4        1.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
5        6.0       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0
6        3.0       5.0  0.0      0.0  …    0.0      0.0     0.0     1.0
7        3.0       NaN  0.0      0.0  …    0.0      1.0     0.0     0.0
8        1.0       NaN  0.0      0.0  …    0.0      0.0     0.0     0.0
9        NaN       0.0  0.0      0.0  …    0.0      0.0     0.0     0.0

   STROKE  TRAUMA  UTI  low_SES  DSFS  length_recoded
0     0.0     0.0  0.0      1.0     8             2.0
1     0.0     0.0  0.0      1.0     4             0.0
2     0.0     0.0  0.0      0.0     5             0.0
3     0.0     0.0  0.0      0.0     5             0.0
4     0.0     0.0  0.0      0.0    11             0.0
5     0.0     0.0  1.0      0.0    10             0.0
6     0.0     0.0  0.0      0.0    20             0.0
7     1.0     0.0  0.0      0.0    13             3.0
8     0.0     0.0  1.0      0.0    22             1.0
9     0.0     0.0  0.0      0.0     4             0.0

[10 rows x 56 columns]
```

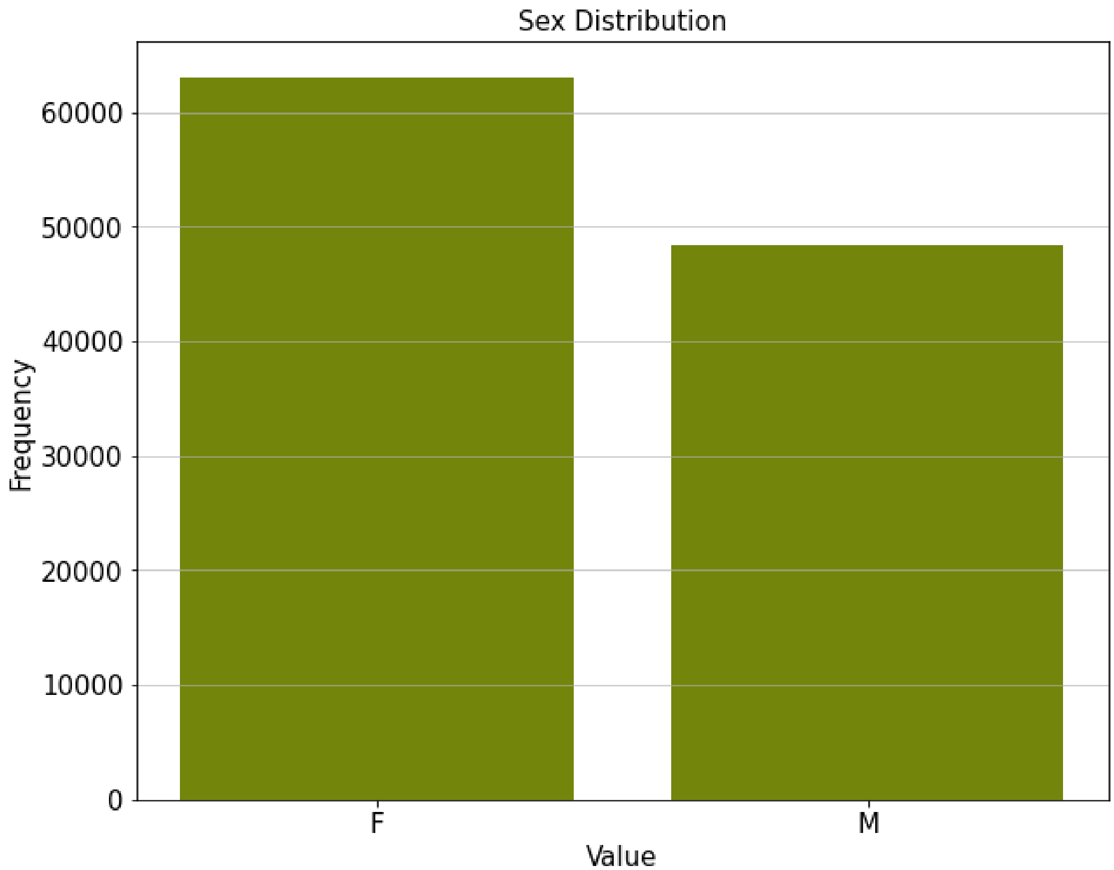
1.4 Exploratory Data Analysis

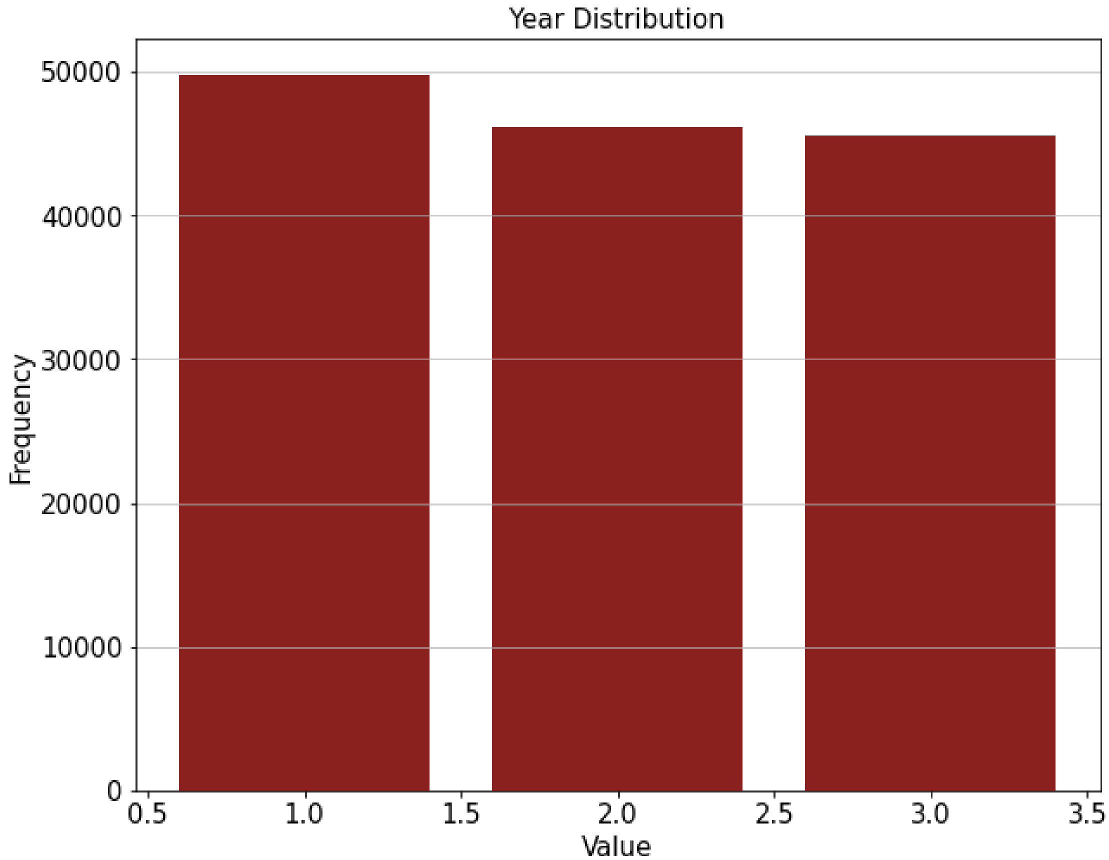
The first step in any data-based problem is understanding the features and outcome we're working with. In addition to visualizing frequencies of specific demographic categories and clinical variables, we'll also visualize the days of hospitalization outcome variable (`length_recoded`).

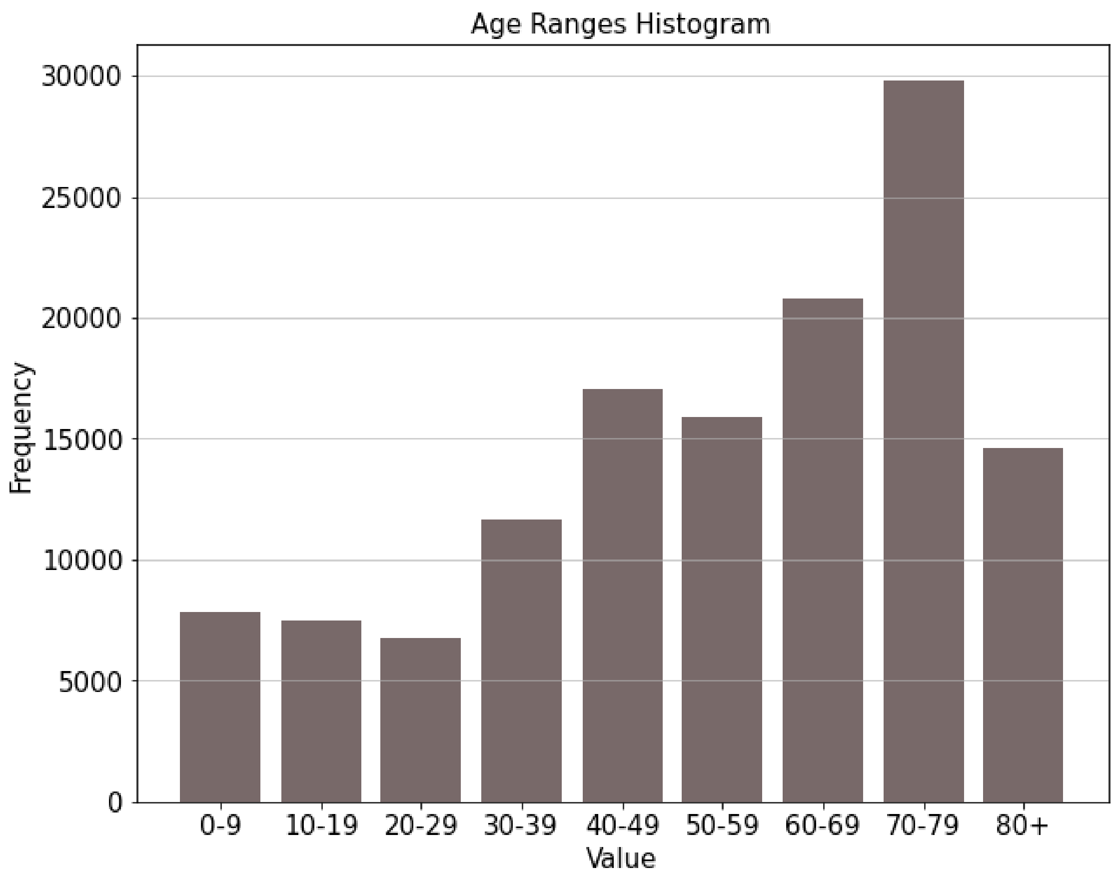
[26]:

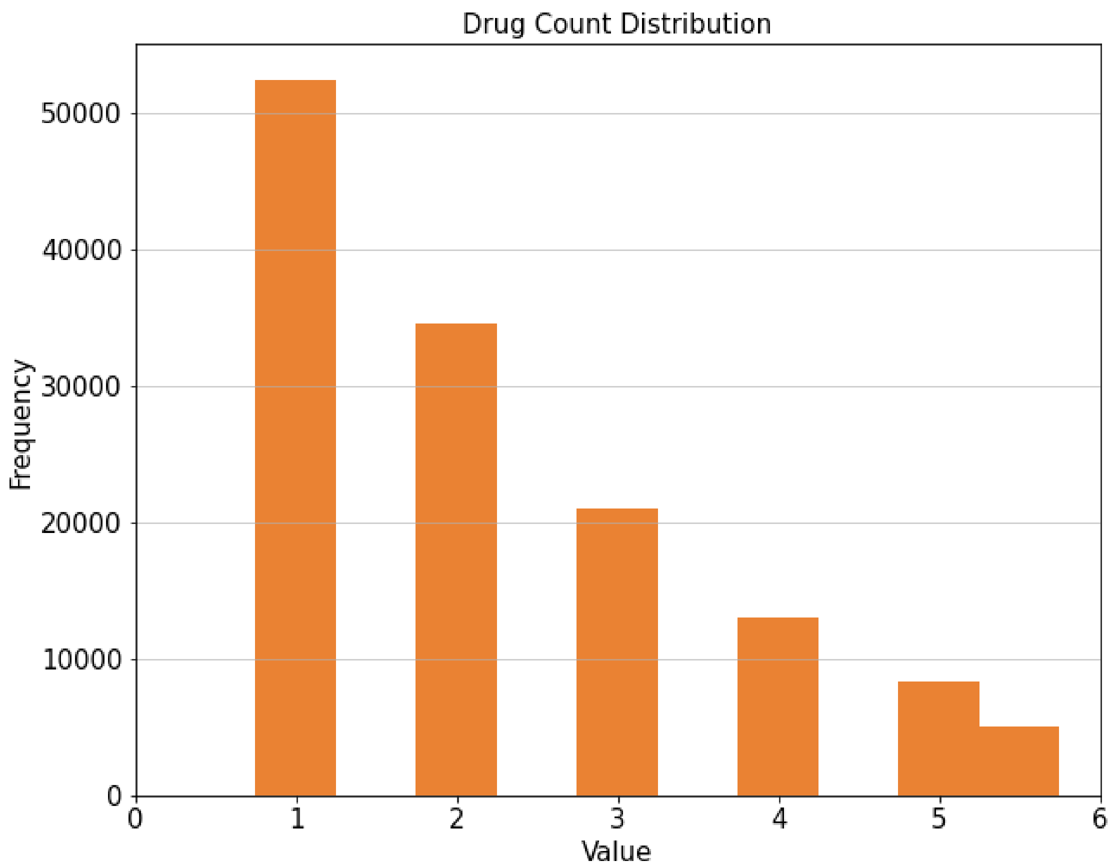
	index	AgeAtFirstClaim
6	0-9	7848
7	10-19	7478
8	20-29	6756
5	30-39	11663
2	40-49	17041
3	50-59	15862

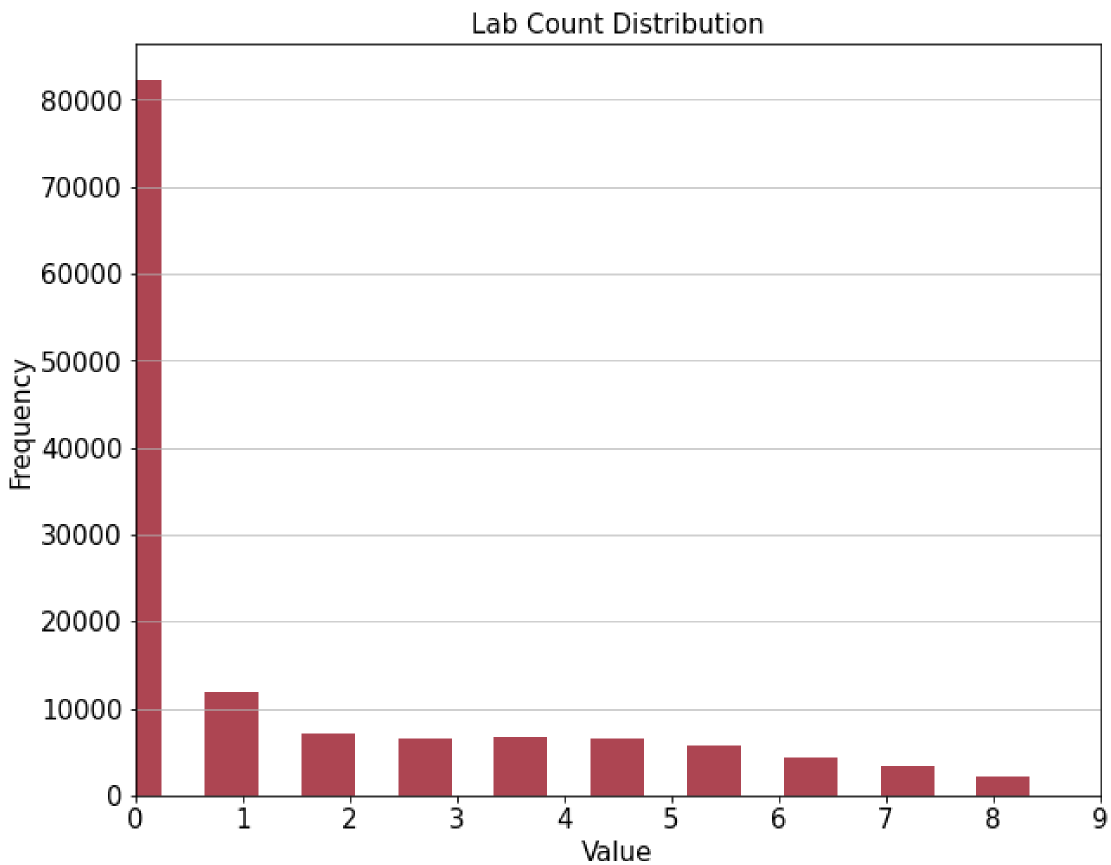
1	60-69	20782
0	70-79	29820
4	80+	14595




Year Distribution
50000
40000
30000
20000
10000
0
Frequency
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Value

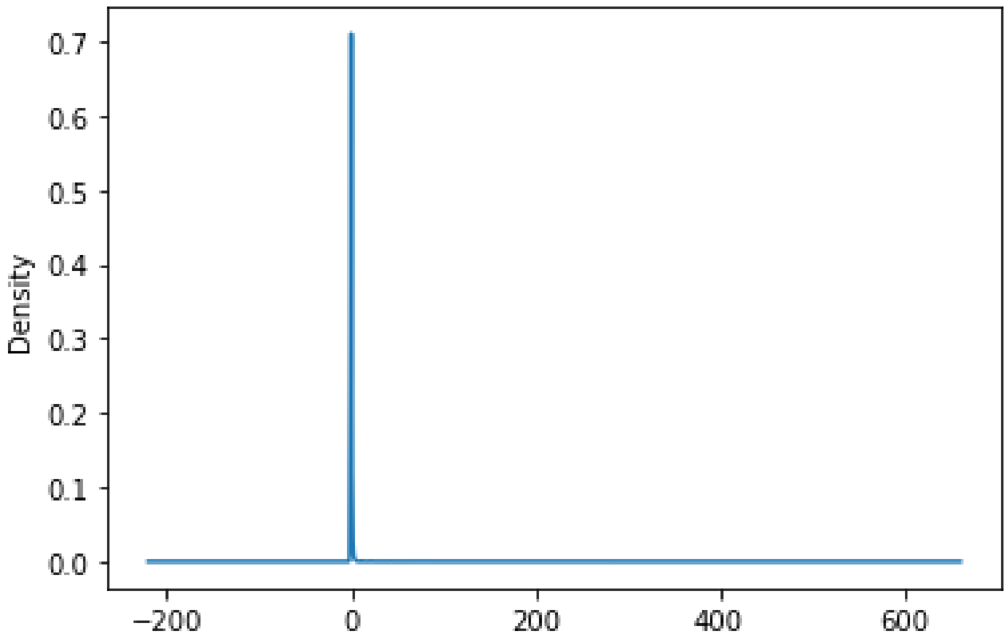

Age Ranges Histogram
Frequency
30000
25000
20000
15000
10000
5000
0
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80+
Value

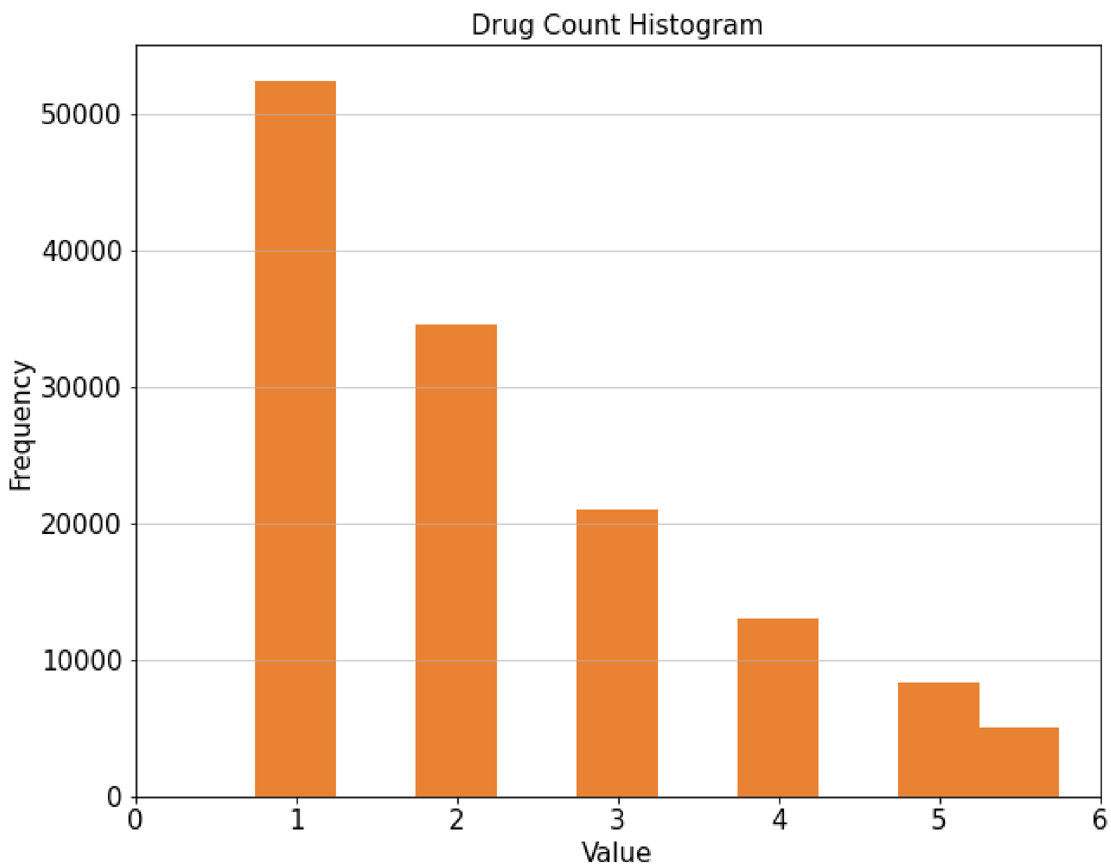

Drug Count Distribution
Frequency
Value
0
10000
20000
30000
40000
50000
0
1
2
3
4
5
6



1.4.1 Outcome Visualized

```
[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7ffcda5db5b0>
```


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
−200
0
200
400
600


Drug Count Histogram
50000
40000
30000
20000
10000
0
Frequency
0
1
2
3
4
5
6
Value

1.5 Modeling

Since we're looking at multiple outcomes with a series of co-variates, decision tree learning would be most appropriate in this scenario. This is further supported knowing that the winning methods used extended decision tree modeling for their predictive analyses.

[128]:

```
   APPCHOL  ARTHSPIN  CANCRA  CANCRB  CANCRM  CATAST  CHF  COPD  FLaELEC  \
0      0.0       0.0     0.0     0.0     0.0     0.0  0.0   0.0      0.0
1      0.0       0.0     0.0     0.0     0.0     0.0  0.0   0.0      0.0
2      0.0       0.0     0.0     0.0     0.0     0.0  0.0   0.0      0.0

   FXDISLC  …  DSFS  Year  AgeAtFirstClaim_10-19  AgeAtFirstClaim_20-29  \
0      0.0  …     8     1                      0                      0
1      0.0  …     4     3                      0                      0
2      0.0  …     5     1                      0                      0

   AgeAtFirstClaim_30-39  AgeAtFirstClaim_40-49  AgeAtFirstClaim_50-59  \
0                      1                      0                      0
1                      1                      0                      0
2                      0                      0                      0

   AgeAtFirstClaim_60-69  AgeAtFirstClaim_70-79  AgeAtFirstClaim_80+
0                      0                      0                    0
1                      0                      0                    0
2                      0                      0                    0

[3 rows x 55 columns]
```

[40]:

```
count    141558.000000
mean          0.342220
std           4.711622
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max         441.000000
Name: length_recoded, dtype: float64
```

[105]:

```
(141558, 55)
```

1.5.1 Linear Regression: Low SES

While `low_SES` was an engineered feature, based on the patient's pay delay, I was curious what it's impact was on other features. This was a question of my own (external to the challenge). I fit an

OLS Linear Regression Model. The `low_SES` coefficient, when controlling for age categories, was statistically significantly associated (= 2.6505, $p < 0.0001$) with the total number of conditions attributed toward hospitalization (per patient). Nothing definitive can be concluded from this, but it is still an interesting observation altogether.

```
                            OLS Regression Results
================================================================================
=======
Dep. Variable:                    sum   R-squared (uncentered):
0.725
Model:                            OLS   Adj. R-squared (uncentered):
0.725
Method:                 Least Squares   F-statistic:
4.672e+04
Date:                Fri, 02 Oct 2020   Prob (F-statistic):
0.00
Time:                        21:02:31   Log-Likelihood:
-3.3906e+05
No. Observations:              141558   AIC:
6.781e+05
Df Residuals:                  141550   BIC:
6.782e+05
Df Model:                           8
Covariance Type:            nonrobust
================================================================================
=========
                           coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---------
AgeAtFirstClaim_20-29     2.5657      0.032     79.290      0.000       2.502
2.629
AgeAtFirstClaim_30-39     2.7627      0.025    111.896      0.000       2.714
2.811
AgeAtFirstClaim_40-49     3.0008      0.021    145.983      0.000       2.960
3.041
AgeAtFirstClaim_50-59     3.3048      0.021    154.550      0.000       3.263
3.347
AgeAtFirstClaim_60-69     3.8142      0.019    199.888      0.000       3.777
3.852
AgeAtFirstClaim_70-79     4.2291      0.017    253.426      0.000       4.196
4.262
AgeAtFirstClaim_80+       4.5627      0.023    197.753      0.000       4.517
4.608
low_SES                   2.6505      0.015    172.416      0.000       2.620
2.681
==============================================================================
Omnibus:                     9157.372   Durbin-Watson:                   1.433
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11675.912
Skew:                           0.612   Prob(JB):                         0.00
Kurtosis:                       3.693   Cond. No.                         2.80
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

1.5.2 Classification Models: Decision Tree

Note that this tree was fit without any dimmensionality reduction. As a result, there's definitely room for pruning and making the model more parsimonious. While the tree is ridiculously large and not as helpful as we'd like, the feature importance is worth noting: besides the time variables, `RESPR4` (*acute respiratory infections*), `ARTHSPIN` (*arthropathies and spine disorders*), `NEUMENT`(*neurological problems*), and `low_SES` were ranked the most important features. Overall, the model was 77.4% accurate.

[50]:
```
DecisionTreeClassifier(random_state=11)
```

[53]:

	features	importance
45	DSFS	0.188571
46	Year	0.062379
1	ARTHSPIN	0.037378
44	low_SES	0.037099
36	RESPR4	0.035005
26	NEUMENT	0.034198
22	MISCHRT	0.031744
18	INFEC4	0.030325
25	MSC2a3	0.027085
40	SKNAUT	0.026664

```
Accuracy: 0.774
```

1.5.3 Classification Models: Random Forest

For this model, I utilized a grid search to optimize parameters based on accuracy and refit accordingly. The model was 86.7% accurate. Furthermore, one of the key advantages of random forest was being able to visualize feature importance. `GIBLEED` and `ROAMI` were the leading clinical features.

```
Accuracy: 0.867
```

[215]:

	name	score
10	GIBLEED	0.15
45	DSFS	0.13
37	ROAMI	0.12

42	TRAUMA	0.09
27	ODaBNCA	0.07